面向 web 图像检索的基于语义迁移的无监督深度哈希 *

许 胜，陈盛双，谢 良

(武汉理工大学 理学院，武汉 430070)

摘 要： 当前主流的 Web 图像检索方法仅考虑了视觉特征，没有充分利用 Web 图像附带的文本信息，并忽略了相关文本中涉及的有价值的语义，从而导致其图像表达能力不强。针对这一问题，提出了一种新的无监督图像哈希方法：基于语义迁移的深度图像哈希（semantic transfer deep visual hashing，STDVH）。该方法首先利用谱聚类挖掘训练文本的语义信息；然后构建深度卷积神经网络将文本语义信息迁移到图像哈希码的学习中；最后在统一框架中训练得到图像的哈希码和哈希函数，在低维汉明空间中完成对大规模 Web 图像数据的有效检索。通过在 Wiki 和 MIR Flickr 这两个公开的 Web 图像集上进行实验，证明了该方法相比其他先进的哈希算法的优越性。

关键词： 语义迁移；图像哈希；Web 图像检索；深度学习

中图分类号： TP391 doi: 10.3969/j.issn.1001-3695.2018.02.0185

Unsupervised deep hashing based on semantic transfer for Web image retrieval

Xu Sheng, Chen Shengshuang, Xie Liang

(*School of Science*, *Wuhan University of Technology*, *Wuhan* 430070, *China*)

Abstract: Most existing Web image retrieval approaches only consider visual features. They ignore the valuable semantics involved in the associated texts, and fail to take advantages of text. This paper proposed a new unsupervised visual hashing approach called semantic transfer deep visual hashing (STDVH) . Firstly, it extracted the semantic information of the training text by spectral clustering. Then, it constructed a deep convolutional neural network to transfer the text semantic information into the learning of the image hash code. At last, it trained the image hash codes and hash functions in a unified framework, and completed the effective retrieval of large-scale image data in low-dimensional Hamming space. Experiments on two publicly available image datasets Wiki and MIR Flickr indicate that the proposed approach can achieve superior performance over other state-of-the-art techniques.

Key words: semantic transfer; image hashing; Web image retrieval; deep learning

0 引言

随着社交媒体和移动计算技术的不断进步，在过去的十年中，Web 图像的可用性得到了巨大的发展。因此，研究智能图像检索技术越来越受到信息检索和多媒体计算领域的关注。特别是基于内容的图像检索（content-based image retrieval，CBIR）[1]，作为仅使用视觉图像作为查询的技术，由于具有广泛的应用前景而变得越来越重要。

为了在海量的图像集合上提供高质量的基于内容的搜索服务，效率和有效性都是亟需研究的重要问题。高效的索引结构对于扩大大数据空间和提高精确搜索至关重要。CBIR 最简单的方法是将查询图像与存储在数据库中的每个样本进行顺序比较。它的线性复杂度导致在实际应用中效率低、可扩展性差。此外，视觉特征通常具有很高的维度。如何解决“维数灾难”仍是一个尚未妥善解决的开放性研究问题。不过在大多数实际的 CBIR 应用中，近似的检索结果可以充分满足用户的信息需求，这表明了近似最近邻检索的可行性。受这种现象的启发，近年来已有多种索引方法被开发出来，如倒排文档[2]、树结构[3]和哈希[4]。倒排文档只能在高维稀疏特征的索引方面表现良好[2]。树结构当被索引的特征的维度变高时，其性能大大降低。而且在存储相应的数据结构时，倒排文档和树结构都会消耗大量的内存。当图像采集规模较大时，这个问题就更加严重。

作为支持快速准确的图像检索的新兴技术之一，图像哈希算法在最近十年得到了极大的关注，成为一个非常活跃的研究领域。其基本思想是将原始高维视觉特征映射为在低维汉明空间中的二进制编码，从而可以通过简单而有效的位操作来衡量图像的视觉相似性。一般来说，图像哈希有两个主要优点：a）快速查询响应，由于按位操作可以被高效地执行，所以可以快

收稿日期： 2018-02-10；**修回日期：** 2018-03-16 **基金项目：** 国家自然科学基金青年基金资助项目（61702388）

作者简介： 许胜（1993-），男，湖北武人，硕士研究生，主要研究方向为深度学习、图像哈希（xusheng556@163.com）；陈盛双（1964-），男，教授，主要研究方向为金融数学、数据挖掘；谢 良（1987-），男，讲师，博士，主要研究方向为多媒体检索、哈希学习.

速完成检索过程；b）低存储消耗，二进制嵌入的结果可以大大降低高维特征的存储。

然而由于视觉特征与人类理解之间存在着“语义鸿沟”，基于视觉特征的哈希会缺失一定的语义信息，从而降低了 CBIR 的性能。为了丰富图像哈希码的语义，已经应用了许多基于机器学习的策略，并提出了多种哈希方案。它们包括无监督图像哈希[5]、有监督图像哈希[6]和半监督图像哈希[7]。有监督和半监督图像哈希都可以提高哈希码的语义判别能力。但是这两种模式在训练过程中都需要标记图像。实际上，这一要求在 CBIR 可能并不能得到满足，这是因为在实际场景中质量高的标记图像很少，而且它们需要大量的人力劳动和专家知识。另一方面，图像（如社交网络中的图片）通常与信息丰富的文本标签或描述相关联。因此有必要考虑利用辅助文本，通过无监督学习来提高图像哈希的质量。而这种方式的核心的挑战是如何开发有效的无监督学习方案，智能地从相关的文本信息中提取和集成语义到图像哈希代码中。

本文提出了一种新的无监督图像哈希方案，称为语义迁移深度图像哈希（semantic transfer deep visual hashing，STDVH）。该方法的关键思想是从图像相关文本中自动提取语义，通过深度学习将语义迁移图像哈希码中，从而提升了图像哈希的性能。STDVH 的工作原理如下：首先通过对文本信息进行谱聚类获取语义，并将其迁移到后续视觉哈希码的学习中；然后构建深度卷积神经网络模型来获取哈希码；最后将语义迁移，哈希码学习与哈希函数的学习整合在统一框架中，从而使得学习到的深度哈希函数能够同时保存原始图像对应的语义信息及其在视觉上的相似性。通过哈希函数能够获取数据库和查询的图像的哈希码，用于图像检索。

本文主要的贡献概括如下：

STDVH 不仅仅只考虑图像特征，或是同时处理图像和文本，而是专门利用文本信息的语义迁移学习来辅助图像哈希。通过谱聚类得到文本语义的聚类，以文本信息聚类结果和图像建立模型，能够将语义有效地结合到哈希码中。

STDVH 采用统一的无监督学习框架，将哈希函数学习，哈希码学习以及语义迁移学习统一在一个框架中，并利用深度卷积神经网络学习原始图像到文本的语义信息。

STDVH 可以利用 Web 图像中包括视觉与文本特征在内的多模态数据进行训练，而且只需要图像的视觉信息作为输入查询。它符合 CBIR 的实际要求，即数据库中的 Web 图像通常附带有文本信息，而用户则不需要提供文本查询。

在公开可用的图像数据库上进行全面的实验。结果充分显示了 STDVH 的优越性，并且证明了 STDVH 从各个方面明显胜过了几种最先进的基于内容或跨模态的哈希方法。

1 相关工作

1.1 单模态图像哈希

根据如何生成哈希函数的方法，现有的单模态图像哈希（SFVH）可以进一步分为数据独立哈希[8]与依赖于数据的哈希[5] 两大类。局部敏感哈希（LSH）[8]是最典型的与数据无关的哈希方法之一，它基于来自例如标准高斯分布等特定分布的随机向量，将相似的点以高概率地映射到同一汉明空间。另一方面，依赖于数据的哈希通过机器学习方法基于底层数据分布的特点来学习哈希函数。谱哈希(SH)[5]是典型的基于无监督学习的哈希方法，通过保留哈希代码中图像的相似性来学习哈希函数。随着哈希技术的发展，稀疏嵌入哈希[9]、基于流形的哈希[10]、基于深度学习的哈希[7]被提出，用来学习有效的二进制哈希码。

1.2 多模态图像哈希

多模态图像哈希中的一种图像形式可以用文本形式来代替。综合多个模态集成对于全面解读图像内容并实现最佳的学习效果非常重要[11]。许多研究人员为了不同的目的而设计各种考虑多特征融合的方案来进行哈希。例如，多视图潜在哈希（MVLH）[11]通过发现多个视图之间共享的潜在因素，将多模态特征结合到二进制表示学习中，根据每个视图的重建误差来学习多特征融合的权重。多视图对齐哈希（MVAH）[12]学习正则化核非负矩阵分解的哈希码，它考虑了多个视觉特征的隐含语义和联合概率分布。

单模态和多模态图像哈希最显着的局限性是它们只考虑了视觉形态的特征。由于语义上的差距，以低级视觉特征为特征的图像关系不能有效地描述丰富的图像语义，从而使得哈希码语义意义较少。

1.3 跨模态哈希

跨模态哈希的核心思想是将异构的模态特征映射到共同的汉明空间，并在该空间中计算相似度来返回跨模态检索结果。Zhou 等人[13]通过采用稀疏编码和矩阵分解，提出了潜在语义稀疏哈希（LSSH）。协同矩阵分解哈希（CMFH）[14]从一个样本的多个模态使用协同矩阵分解与潜在因子模型学习哈希码。

由于跨模态哈希的投影空间可能嵌入比单模态视觉特征空间更多的语义，跨模态哈希可以提高 CBIR 的性能。但是各种跨模态哈希方法的主要设计目标是在不同的模式下进行多媒体检索。它假定每种涉及的模态都对跨模态检索有同样的贡献。这个假设使得它们共享相同的汉明空间，所以并能够对专门的图像视觉信息进行有效判别。另外，原始图像视觉特征中特有的判别信息由于强制进行异构模态的关联，而可能导致在哈希过程中丢失特色语义信息。

表 1 总结了最先进的哈希方法和本文提出的 STDVH 的关键特征。基于上面的分析，可以发现专门设计一个智能哈希方法来有效地利用相关的模态（例如文本信息等）来辅助图像哈希是非常重要的。

表 1 主要的无监督图像哈希和 STDVH 的特征

方法	查询	学习特征	学习空间	语义增强	CBIR
单模态图像哈希	视觉	视觉	视觉	否	是
多模态图像哈希	视觉	视觉	视觉	否	是

多模态哈希	视觉+文本	视觉+文本	多模态	是	否
无监督跨模态哈希	视觉/文本	视觉+文本	共享	有限的	部分
STDVH	视觉	视觉+文本	文本增强	是	是

2　基于语义迁移的无监督深度哈希

2.1　系统概述

图 1 描述基于语义迁移的无监督深度哈希运用于图像检索的系统框架。系统主要包含离线学习和在线检索两个部分。

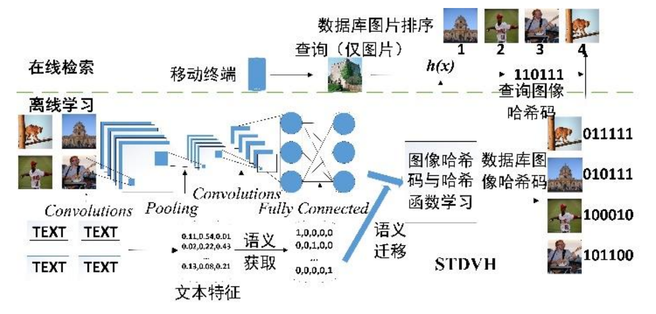


图 1　基于 STDVH 的 CBIR 系统框架

a)离线学习。该部分的目的是从数据库图像中学习哈希码，同时生成查询图像的哈希函数。其包含三个主要步骤：首先，对训练集中的文本特征进行谱聚类，利用文本的谱聚类结果来增强图像哈希码的学习效果；然后，以训练集的原始图像为输入，以文本信息谱聚类的结果为输出构建卷积神经网络，训练整个网络，得到卷积神经网络的倒数第三层输出转化为哈希码；最后将语义迁移，哈希码学习与哈希函数的学习整合在统一框架中，从而使得学习到的深度哈希函数能够同时保存原始图像对应的语义信息及其在视觉上的相似性。

b)在线检索。提取查询图像，利用哈希函数将其映射成二进制码。最后计算查询图像和数据库图像之间的汉明距离，并以距离由小到大的顺序返回数据库图像。

2.2　符号与问题描述

本文使用粗体大写字母来表示矩阵，使用粗体小写字母来表示向量。矩阵 $\mathbf{X}$ 的转置表示为 $\mathbf{X}^{\mathrm{T}}$，矩阵 $\mathbf{X}$ 的逆表示为 $\mathbf{X}^{-1}$，$tr(\mathbf{X})$ 表示矩阵 $\mathbf{X}$ 的迹，$\|\cdot\|_F$ 表示弗罗贝尼乌斯范数。$\mathrm{sgn}(\cdot)$ 表示符号函数，如果是正则返回 1，否则返回-1。

设有 N 对数据点 $\mathbf{X}^{(m)}=\left\{\mathbf{x}_i^{(m)}\right\}_{i=1}^{N}, m=1,2$，其中 $\mathbf{x}_i^{(1)}$ 表示第 i 张图片，$\mathbf{x}_i^{(2)}$ 表示对应的文本特征。无监督图像哈希的目标是学习数据库图像 $\mathbf{X}^{(1)}$ 的哈希码 $\mathbf{Y}=\{\mathbf{y}_i\}_1^N\in\mathbb{R}^{L\times N}$，其中 $\mathbf{y}_i\{-1,1\}^L$ 是第 i 张图片的哈希码，二值哈希码可以写做 $\mathbf{y}_i=h(\mathbf{x}_i^{(1)})=[h_1(\mathbf{x}_i^{(1)}),h_2(\mathbf{x}_i^{(1)}),\cdots,h_L(\mathbf{x}_i^{(1)})]^{\mathrm{T}}$，其中 $h(\mathbf{x}_i^{(1)})$ 是需要学习的哈希函数，L 是图像哈希码的长度。

2.3　语义学习

本文将文本特征的语义信息迁移到视觉哈希码的学习中，以此来增强视觉哈希的效果。在 STDVH 模型中，考虑到谱聚类能够在任意形状的样本空间上聚类，而且收敛于全局最优，本文使用谱聚类来产生文本特征的语义信息。确切地说，对于训练集中的文本特征 $\mathbf{X}^{(2)}=\left\{\mathbf{x}_i^{(2)}\right\}_{i=1}^{N}$，通过谱聚类产生类别矩阵 $\mathbf{C}=\{\mathbf{c}_i\}_{i=1}^{N}$，其中 $\mathbf{c}_i$ 表示第 i 个数据点的文本特征对应的类别，即语义信息。然后根据类别矩阵 $\mathbf{C}$ 生成文本相似度量矩阵 $\mathbf{S}=\{s_{ij}\}$，将该相似矩阵迁移到视觉哈希的学习中。

对于训练集中的文本特征 $\mathbf{X}^{(2)}=\left\{\mathbf{x}_i^{(2)}\right\}_{i=1}^{N}$，将其分成 K 个类。根据数据构建图 $\mathbf{G}=(\mathbf{X}^{(2)},\mathbf{E})$，顶点表示各个文本信息，带权的边表示文本之间的相似度。

设 $\mathbf{A}_1,\mathbf{A}_2,\cdots,\mathbf{A}_K$ 为图的几个子集，这些子集没有交集，为了让分割的 Cut 最小，谱聚类便是要最小化下述目标函数：

$$cut(\mathbf{A}_1,\mathbf{A}_2,\cdots,\mathbf{A}_K)=\frac{1}{2}\sum_{i=1}^{K}\mathbf{T}(\mathbf{A}_i,\bar{\mathbf{A}}_i),\tag{1}$$

其中：$\mathbf{A}_i$ 表示第 i 个组；$\bar{\mathbf{A}}_i$ 表示 $\mathbf{A}_i$ 的补集；$\mathbf{T}(\mathbf{A}_i,\bar{\mathbf{A}}_i)$ 表示 $\mathbf{A}_i$ 组与 $\bar{\mathbf{A}}_i$ 之间所有边的权重之和。

用邻接矩阵 $\mathbf{M}=\left\{m_{ij}\,|\,1\le i\le N,1\le j\le N\right\}$ 表示图，其相似性是按照式（2）计算。

$$m_{ij}=e^{\frac{-cos(\mathbf{x}_i^{(2)},\mathbf{x}_j^{(2)})^2}{2\sigma^2}},\tag{2}$$

其中：σ 是一个超参数；$cos(\mathbf{x}_i^{(2)},\mathbf{x}_j^{(2)})$ 表示余弦距离。为了排除自身相似度的影响，将对角线上的元素赋值为 0，即 $m_{ii}=0$。

为了使谱聚类效果更好，将相似度矩阵 $\mathbf{M}$ 按照式（3）稀疏化。

$$\mathbf{M}(m_{ij}<\lambda\bar{m})=0,\tag{3}$$

其中：λ 为稀疏度；$\bar{m}$ 表示所有元素的平均值。

为了使某个单节点不会更容易被剔除，本文考虑一个归一化的对角矩阵 $\mathbf{D}$，对角线上元素是相似度矩阵一行（列，因为对称行列一样）所有元素的和，即

$$\mathbf{D}(i,i)=\sum_{j=1}^{N}m_{ij}.\tag{4}$$

然后计算归一化拉普拉斯图矩阵 $\mathbf{L}$：

$$\mathbf{L}=\mathbf{D}^{1/2}\mathbf{M}\mathbf{D}^{1/2}.\tag{5}$$

计算拉普拉斯图矩阵 $\mathbf{L}$ 的特征值和特征向量，将特征值从大到小排列，选取前 K 个特征值对应的特征向量，将其组合成一个矩阵 $\mathbf{V}$，即 $\mathbf{V}=\left[\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_K\right]$，其中 $\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_K$ 为前 K 个特征值最大的特征向量。

对矩阵 $\mathbf{V}$ 中的每一行进行单位化处理，得到矩阵 $\mathbf{P}$，即

$$\mathbf{P}_{ij}=\frac{\mathbf{V}_{ij}}{(\sum_{j=1}^{N}\mathbf{V}_{ij}^{\;2})^{1/2}}.\tag{6}$$

把矩阵 $\mathbf{P}$ 的每一行看成 K 维空间中的点，利用传统的聚类算法，这里采用 K-means 算法[15]，将其聚成 K 类。聚类的结果中每一行所属于的类别就是原来文本特征分别所属的类别，得到类别矩阵 $\mathbf{C}=\{\mathbf{c}_i\}_{i=1}^{N}$。语义学习的过程归纳为算法 1。

算法 1　语义学习

输入：

训练集文本特征 $\mathbf{X}^{(2)}=\{\mathbf{x}_i^{(2)}\}_{i=1}^{N}$，聚类数 K，超参数 σ，稀疏度 λ。

输出：

语义类别矩阵 $\mathbf{C}=\{\mathbf{c}_i\}_{i=1}^{N}$。

1：计算相似度矩阵 $\mathbf{M}$

2：将 $\mathbf{M}$ 的对角线值赋值 0，$m_{ii}=0$

3：按公式(3)稀疏化 $\mathbf{M}$

4：计算归一化矩阵 $\mathbf{D}$

5：计算归一化拉普拉斯图矩阵 $\mathbf{L}$

6：计算 $\mathbf{L}$ 的特征向量，将前 K 个特征值最大的向量按列组合成一个矩阵 $\mathbf{V}$，即 $\mathbf{V}=[\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_K]$，$\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_K$ 为前 K 个特征值最大的特征向量

7：归一化 $\mathbf{V}$ 形成矩阵 $\mathbf{P}$

8：对矩阵 $\mathbf{P}$ 按每一行为数据点，进行 K-means 聚类得到 $\mathbf{C}$

2.4　无监督深度哈希

本文的 STDVH 模型包含一个 CNN 模型。无监督深度哈希是一个拥有 9 层的卷积神经网络，其中前面 7 层和 AlexNet[16]一样。当然，其他的 CNN 结构也可以取代 AlexNet 在 STDVH 中的作用，但是本文的目的不是研究不同的网络。所以这里只用 AlexNet 作为 STDVH 模型中深度哈希的一部分，对于其他网络将来再作研究。基于前面的语义学习结果，将原始图像作为 CNN 的输入，对应的语义类别作为输出。

表 2 展示了 STDVH 深度哈希部分的详细配置。该部分包含 5 个卷积层（conv 1-5）和 4 个全连接层（full 6-9）。每个卷积层从以下几个方面描述：“filter”表示卷积滤波器的数目和它们接收域的尺寸以及通道数，形如“数目　尺寸 x 尺寸 x 通道”；“stride”表示卷积步幅，即将滤波器应用于输入的间隔；“pad”表示要添加到输入的每一侧的像素的数量;“LRN”表示是否应用局部影响归一化层（local response normalization，LRN）；“pool”表示下采样因子；全连接层中的“4096”表示输出的维度；“L”表示哈希码的长度；“K”表示文本语义产生的类别数。所有层的激活函数是线性整流函数（REctification Linear Unit，RELU）。

表 2　无监督深度哈希的网络参数

Layer	Configuration
conv1	filter 96 11x11x3, stride 4x4, pad 0, LRN, pool 2x2
conv2	filter 256 5x5x48, stride 1x1, pad 2, LRN, pool 2x2
conv3	filter 384 3x3x256, stride 1x1, pad 1
conv4	filter 384 3x3x192, stride 1x1, pad 1
conv5	filter 256 3x3x192, stride 1x1, pad 1, pool 2x2
full6	4096
full7	4096
full8	L
full9	K

2.5　基于统一框架的哈希学习

本文搭建一个统一的无监督框架将语义迁移，哈希码学习以及哈希函数学习整合在一起。

根据语义学习的结果 $\mathbf{C}$，可以得到数据点的相似性矩阵 $\mathbf{S}=\{s_{ij}\}$，其中 $s_{ij}\in\{0,1\}$。当 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 相似时 $s_{ij}=1$，$\mathbf{x}_i$ 和 $\mathbf{x}_j$ 不相似时 $s_{ij}=0$。定义如下：

$$s_{ij}=\begin{cases}1 & \mathbf{c}_i=\mathbf{c}_j,\\ 0 & \text{otherwise.}\end{cases} \tag{7}$$

对于给出的所有图片的二值码 $\mathbf{Y}=\{\mathbf{y}_i\}_1^N\in\mathbb{R}^{L\times N}$，定义对于相似矩阵 $\mathbf{S}=\{s_{ij}\}$ 的最大似然估计如下：

$$p(s_{ij}\mid\mathbf{Y})=\begin{cases}\sigma(\Omega_{ij}) & s_{ij}=1,\\ 1-\sigma(\Omega_{ij}) & s_{ij}=0,\end{cases} \tag{8}$$

其中：$\Omega_{ij}=\frac{1}{2}\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j$；$\sigma(\Omega_{ij})=\frac{1}{1+e^{-\Omega_{ij}}}$。这里要注意的是 $\mathbf{y}_i\in\{-1,1\}^L$。

通过采取负对数似然，可以得到以下优化问题：

$$\begin{aligned}\min_{\mathbf{Y}}\tau_1&=-\log p(\mathbf{S}\mid\mathbf{Y})=-\sum_{s_{ij}\in\mathbf{S}}\log p(s_{ij}\mid\mathbf{Y})\\&=-\sum_{s_{ij}\in\mathbf{S}}(s_{ij}\Omega_{ij}-\log(1+e^{\Omega_{ij}})).\end{aligned} \tag{9}$$

显然，上面的优化问题式（9）可以使得两个相似点之间的汉明距离尽可能地小，同时使得两个不相似点之间的汉明距离尽可能地大，这符合无监督图像哈希的目的。

本文以一种离散的方式来解决这个问题，将其转换成以下等价形式：

$$\begin{aligned}&\min_{\mathbf{Y},\mathbf{U}}\tau_2=-\sum_{s_{ij}\in\mathbf{S}}(s_{ij}Z_{ij}-\log(1+e^{Z_{ij}}))\\&s.t.\quad \mathbf{u}_i=\mathbf{y}_i,\mathbf{u}_i\in\mathbb{R}^{L\times1},\mathbf{y}_i\in\{-1,1\}^L,\forall i=1,2,\cdots,N.\end{aligned} \tag{10}$$

其中：$Z_{ij}=\frac{1}{2}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j$；$\mathbf{U}=\{\mathbf{u}_i\}_{i=1}^{N}$；$\mathbf{u}_i$ 是二值哈希码 $\mathbf{y}_i$ 在松弛条件下的连续状态表达，利用式子 $\mathbf{y}_i=\mathrm{sgn}(\mathbf{u}_i)$ 便可得到离散化的哈希码。

为了优化问题式(10)，可以通过将式(10)中的等式约束移到正则化项来优化正则化问题。

$$\begin{aligned}\min_{\mathbf{Y},\mathbf{U}}\tau_3=&-\sum_{s_{ij}\in\mathbf{S}}(s_{ij}Z_{ij}-\log(1+e^{Z_{ij}}))\\&+\eta\sum_{i=1}^{N}\|\mathbf{y}_i-\mathbf{u}_i\|_2^2,\end{aligned} \tag{11}$$

其中：η 是正则化项。

传统的哈希方法通常依赖于手工特征提取，而且哈希学习

阶段与特征提取是分离的，造成提取的特征并不能与哈希过程最优适配，这些特征往往不能保持语义上的相似性。为了能使图像的特征表示和哈希码可以互相促进提升，将深度哈希与目标函数对应起来，令

$$\mathbf{u}_i=\mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i^{(1)};\theta)+\mathbf{v}, \tag{12}$$

其中：θ 表示深度哈希部分的 CNN 网络前面 7 层的所有参数；$\phi(\mathbf{x}_i^{(1)};\theta)$ 表示图片 $\mathbf{x}_i^{(1)}$ 的第 7 层的输出；$\mathbf{W}\in\mathbb{R}^{4096\times L}$ 是一个权值矩阵；$\mathbf{v}\in\mathbb{R}^{L\times 1}$ 是偏移向量。这表示在深度哈希部分用一个权值矩阵 $\mathbf{W}$ 和偏移向量 $\mathbf{v}$ 的全连接层将图像特征部分和语义迁移结果连接起来。那么问题表征如下：

$$\begin{aligned}\min_{\mathbf{Y},\mathbf{W},\mathbf{v},\theta}\tau=&-\sum_{s_{ij}\in\mathbf{S}}(s_{ij}Z_{ij}-\log(1+e^{Z_{ij}}))\\&+\eta\sum_{i=1}^{N}\left\|\mathbf{y}_i-(\mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i^{(1)};\theta)+\mathbf{v})\right\|_2^2.\end{aligned} \tag{13}$$

这样，就将语义迁移、哈希函数学习以及哈希码学习整合同一个框架中。

在 STDVH 模型中，学习部分的参数包含 $\mathbf{W},\mathbf{v},\theta$ 和 $\mathbf{Y}$。采用小批量的学习策略，在每次迭代中，从整个训练集中采样小部分数据点，然后根据这些采样点进行学习。使用交替学习的优化方法，用固定的其他参数来优化一个参数。

对于 $\mathbf{y}_i$，直接优化如下：

$$\mathbf{y}_i=\mathrm{sgn}(\mathbf{u}_i)=\mathrm{sgn}(\mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i^{(1)};\theta)+\mathbf{v}). \tag{14}$$

对于其他参数 $\mathbf{W},\mathbf{v}$ 和 θ，采用反向传播算法进行学习，按照式（15）计算关于 $\mathbf{u}_i$ 的损失函数的导数。

$$\begin{aligned}\frac{\partial\tau}{\partial\mathbf{u}_i}=&\frac{1}{2}\sum_{j:s_{ij}\in\mathbf{S}}(a_{ij}-s_{ij})\mathbf{u}_j+\frac{1}{2}\sum_{j:s_{ji}\in\mathbf{S}}(a_{ji}-s_{ji})\mathbf{u}_j\\&+2\eta(\mathbf{u}_i-\mathbf{y}_i),\end{aligned} \tag{15}$$

其中：$a_{ij}=\sigma(\frac{1}{2}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j)$。

然后，利用反向传播分别更新参数 $\mathbf{W},\mathbf{v}$ 和 θ：

$$\frac{\partial\tau}{\partial\mathbf{W}}=\phi(\mathbf{x}_i^{(1)};\theta)(\frac{\partial\tau}{\partial\mathbf{u}_i})^{\mathrm{T}}, \tag{16}$$

$$\frac{\partial\tau}{\partial\mathbf{v}}=\frac{\partial\tau}{\partial\mathbf{u}_i}, \tag{17}$$

$$\frac{\partial\tau}{\phi(\mathbf{x}_i^{(1)};\theta)}=\mathbf{W}\frac{\partial\tau}{\partial\mathbf{u}_i}. \tag{18}$$

算法 2　深度图像哈希学习

输入：

　训练集图像 $\mathbf{X}^{(1)}=\{\mathbf{x}_i^{(1)}\}_{i=1}^{N}$，语义相似度 $\mathbf{S}=\{s_{ij}\}$。

输出：

　参数 $\mathbf{W},\mathbf{v},\theta$ 和 $\mathbf{Y}$。

初始化：初始深度哈希模型 CNN 网络的参数 θ，通过从均值为 0 和方差为 0.01 的高斯分布随机采样来初始化 $\mathbf{W}$ 和 $\mathbf{v}$ 的每一项。

重复

　从 $\mathbf{X}^{(1)}$ 中随机抽样小批量点，并对每个采样点 $\mathbf{x}_i^{(1)}$ 执行以下操作：

　1：通过向前传播计算 $\phi(\mathbf{x}_i^{(1)};\theta)$；

　2：计算 $\mathbf{u}_i=\mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i^{(1)};\theta)+\mathbf{v}$；

　3：利用 $\mathbf{y}_i=\mathrm{sgn}(\mathbf{u}_i)$ 计算 $\mathbf{x}_i^{(1)}$ 的二值码；

　4：通过公式(16)(17)(18)计算图像 $\mathbf{x}_i^{(1)}$ 的偏导；

　5：利用反向传播更新参数 $\mathbf{W},\mathbf{v}$ 和 θ；

结束 固定的迭代次数

哈希函数学习部分归纳为算法 2。完成学习过程后，只能得到训练数据中的点的哈希码。仍然需要执行样本外扩展来预测未出现在训练集中的点的哈希码。STDVH 的深度哈希框架可以自然地扩展到样本外。对于任何图像 $\mathbf{x}_q\notin\mathbf{X}^{(1)}$，使用下面的哈希函数来预测它的哈希码：

$$\mathbf{y}_q=h(\mathbf{x}_q)=\mathrm{sgn}(\mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_q;\theta)+\mathbf{v}). \tag{19}$$

3　实验构造

3.1　实验数据集

为了验证基于语义迁移的无监督深度哈希的有效性，本文在 Wiki[17]和 MIR Flicker[18]这两个公开可用的图像数据集上进行了综合实验。所有的数据集都是由图像和文本对组成，在过去的工作中被广泛用于评价多媒体检索的性能。在相同的设置下，所有的数据集都被划分为查询集、学习集和数据库集。这个实验设置是符合 CBIR 的实际应用场景。

表 3　实验数据统计

数据集	Wiki	MIR Flickr
数据大小	2866	25000
查询大小	1000	1500
训练大小	1500	2000
视觉形式	原始图像	原始图像
文本形式	文本主题	文本词袋

Wiki 包含了 10 种语义类别的 2 866 对多媒体文档，这些数据集从维基百科上搜集得到的。视觉内容以原始图像表示，文本内容用通过潜在狄利克雷分配生成的 10 维主题向量表示。对于 Wiki 数据集，由于图片已经标记成 10 种不同的类别，该数据集中的图像只有在同一类别时才认为是相关的。

MIR Flickr 包含了从 Flickr 得到的 38 个类别的 25 000 张图像。每张图像都有文本标签。为了排除与图像内容不相关的标签的影响，将出现少于 50 次的文本标签删除，这样总共产生了 457 个标签的词汇表[19]。视觉内容以原始图像表示，文本内容以 457 维二元向量表示。每一维都表示了对应标签是否附属于该图像。由于 MIR Flickr 中的图像通常属于多个类别，所以只有当它们至少属于一个共同类别的时候才认为是相关的。

3.2　评价标准

本文实验研究中，平均查准率均值(mAP)被采用为评价指标[14]。对于一个给定的查询，计算平均精度(AP)根据公式 $AP=\frac{1}{NR}\sum_{r=1}^{R}\psi(r)\varphi(r)$。其中：$R$ 是返回结果的个数；NR 是返回结果中的相关图像的数量；$\psi(r)$ 表示的精度最高 r 检索图像，它被定义为图像检索的相关图片和图像 r 的相关查询的数值比

例；$\varphi(r)$ 是指标函数，如果第 r 张图片是相关查询就等于 1，反之等于 0。mAP 被定义为所有查询的 AP 的平均值，其值越大意味着检索性能更好。在实验中，将 R 设为 50 来获取结果。此外，还给出了 Precision-Scope 曲线以反映检索性能相对于检索图像的数量的变化。

3.3　比较方法

该方法是专门为 CBIR 设计的，没有使用任何有监督信息的图像。因此，为了比较公平，本文比较有一些最先进的单模态与跨（多）模态哈希方法。用于对比的单模态哈希方法有迭代量化（ITQ）[20]、局部敏感哈希（LSH）[8]、PCA 哈希（PCAH）[21]、谱哈希（SH）[5]、随机旋转 PCA 哈希（PCA-RR）[20]和密度敏感哈希(DSH)[22]。用于对比的无监督跨模态哈希包括典型相关分析哈希（CCA）[20]、协同矩阵分解哈希（CMFH）[14]。其中 CMFH 通过多种模式共享学习一个潜在的语义子空间，视觉和文本特征都被映射成一个统一的哈希码。

需要注意的是，CCA 和 CMFH 可以为查询视觉图像和文本生成哈希码。实验的目的是为了测试其性能 CBIR，因此本文去掉了文本的哈希码。在这种情况下,所有比较方法的 CBIR 的检索过程均基于视觉哈希码的汉明距离。所有参数的比较方法是根据相关文献和报告进行调整最佳的性能。

3.4　实现细节

在实验中，通过多次对比实验选择参数。对于 Wiki 数据集，选取了谱聚类数量 $K=10$，相似度矩阵中超参数 $\sigma=0.1$，相似度矩阵的稀疏度 $\lambda=1$，这种条件下性能最佳。对于 MIR Flickr 数据集，$K=15$，$\sigma=1$，由于文本数据为二元向量，其相似度矩阵足够稀疏，在提取语义时无需稀疏化，所以选取稀疏度 $\lambda=0$。在实验中，为了观察性能所有数据集上的哈希码长度 L 的范围是[16,32,64,128]。检索范围设置为 100~1000，步长是 100。

在深度哈希部分，首先将所有图像的大小调整为 227×227 像素；然后直接使用原始图像作为输入，将对文本特征进行谱聚类产生的语义信息作为输出。采用已经在 ImageNet[16]数据集上预先训练的 AlexNet 网络来初始化 STDVH 框架的前 7 层。

4　结果与讨论

STDVH 和所有比较的方法在不同数据集上的不同哈希码长度的 mAP 的结果如表 4 所示。两个数据集上 128 bit 的 Precision-Scope 曲线如图 2 所示。根据所得到的结果，可以清楚地看到 STDVH 超越了所有比较方法。随着哈希码长度的增加，STDVH 的检索性能稳定增强，然而对于其他一些被比较的方法来说，哈希码的检索性能随长度增加的改善并不明显，说明 STDVH 学习到的哈希码具有较少的信息冗余。此外，在使用较少的哈希码位数时，STDVH 可以获取比其他使用更长位数的方法更好的性能。原因在于，在文本语义迁移的帮助下，STDVH 可以将更多的语义信息压缩成短哈希码。这意味着基于 STDVH 的 CBIR 可以在相同的性能水平下拥有更快的检索过程和更低的存储成本。

表 4　用于比较的无监督哈希方法的 mAP

方法	Wiki				MIR Flickr			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128bit
ITQ	0.2088	0.2067	0.2151	0.2233	0.6808	0.6993	0.7027	0.7157
LSH	0.2012	0.2061	0.2112	0.2143	0.6603	0.6755	0.7025	0.7100
PCAH	0.2169	0.2164	0.2141	0.2039	0.6800	0.6780	0.6825	0.6865
SH	0.2002	0.2090	0.1968	0.2130	0.6756	0.6819	0.6767	0.6734
PCA-RR	0.2107	0.2085	0.2085	0.2177	0.6758	0.6842	0.7042	0.7176
DSH	0.2107	0.2096	0.2087	0.2219	0.6609	0.6713	0.6921	0.7072
CCA	0.2078	0.2019	0.1959	0.1932	0.5909	0.6004	0.6147	0.6322
CMFH	0.2209	0.2228	0.2321	0.2319	0.6821	0.6909	0.7121	0.7171
STDVH	**0.3370**	**0.3421**	**0.3518**	**0.3632**	**0.6990**	**0.7182**	**0.7460**	**0.7575**

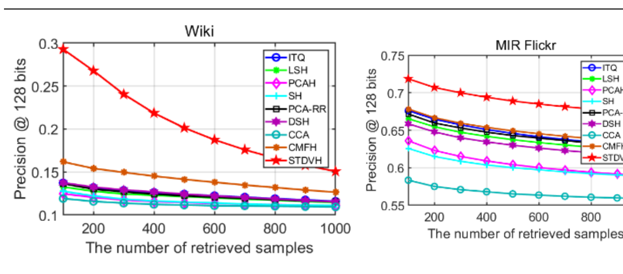


图 2　128bit 哈希码长度的 Precision-Scope 曲线

4.1　语义迁移对视觉哈希的影响

本文通过实验来验证文本语义迁移对提高视觉哈希语义有效性，分别以文本特征和图像特征的谱聚类结果来搭建深度视觉哈希模型，将 STDVH 的性能与忽略文本信息的区别仅考虑视觉特征的性能进行比较。图 3 给出了详细的实验结果。可以看出，语义迁移可以提高 CBIR 的检索性能。其表现更好的原因是，在语义的帮助下，图像与图像和潜在共同主题之间的关系可以更好地进行建模和关联。提取的有价值的语义可以在二进制哈希码中有效地编码。在不同的数据集和哈希码长度上的性能差距是不同的，主要是由于文本对视觉哈希的辅助效果不同造成的。

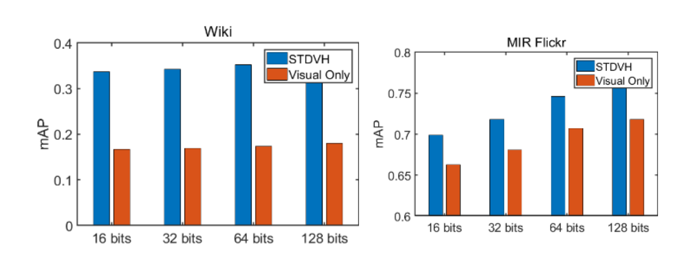


图 3　两个数据集上语义迁移的影响

4.2　训练集大小的影响

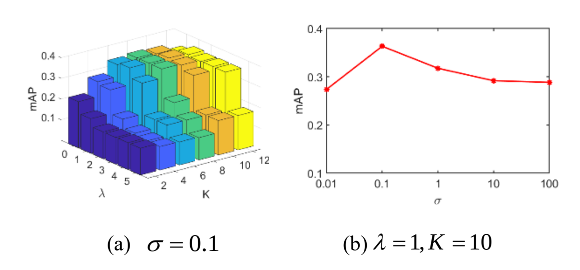
本节构造实验来观察 MIR Flickr 上训练集大小的性能变化。将哈希码的长度设定为 128 bit，并在训练集大小从 1 000 变为 10 000 时记录性能变化。表 5 显示了主要结果。当更多的训练数据被利用时，STDVH 的 mAP 略有增加，说明了 STDVH 学习哈希函数的稳定性。通过进一步观察验证了，在训练数据有限的情况下，文本的语义迁移可以有效地缓解视觉哈希码的语义不足。

表 5　MIR Flickr 关于训练集大小的性能变化

训练集大小	1K	2K	3K	4K	5K
STDVH	0.7454	0.7575	0.7671	0.7716	0.7831
训练集大小	6K	7K	8K	9K	10K
STDVH	0.7872	0.7900	0.7973	0.7999	0.8012

4.3　参数灵敏度

本节通过实验来观察 STDVH 中参数对性能变化的影响。K 表示文本信息谱聚类中的聚类类别数，σ 为超参数，λ 为相似度矩阵的稀疏度。将哈希码的长度设定为 128 bit，并在 Wiki 数据集上进行实验。测试参数 K 从 2 到 12 变化，参数 σ 从数量级（0.01,0.1,1,10,100）的变化，参数 λ 从 0 到 5 的范围变化。在实验中，固定一个参数并观察剩下两个参数的变化。详细的实验结果如图 4 所示。从图 4（a）可以看出，当聚类数 $K=10$，稀疏度 $\lambda=1$ 时性能相对要好很多；从（b）可以看出，当 σ 到达某一点即 $\sigma=1$ 时性能最佳。



(a)　$\sigma=0.1$　　　(b) $\lambda=1, K=10$

图 4　Wiki 数据集上 STDVH 的参数 K,σ,λ 的性能变化

5　结束语

大多数现有针对 CBIR 的哈希方法只考虑了视觉特征。它们忽略了相关文本中涉及的有价值的语义。本研究提出了一个有效的哈希框架 STDVH。利用图像的相关文本提取语义迁移到无监督视觉哈希的学习中。构建深度卷积神经网络将额外的判别语义信息整合到视觉哈希代码和函数中，同时也保留了图像视觉的相似性。离线学习可以有效地利用文本中涉及的语义，而在线哈希只需要视觉图像作为输入，符合 CBIR 实际应用场景的要求。在几个标准图像数据集上进行综合实验，验证了在文本的辅助下视觉哈希的性能可以得到提高，与一些现有的哈希技术相比，STDVH 有着更好的性能。

参考文献：

[1] Su Jahwung, Huang Weijyun, Yu Philip S, *et al.* Efficient relevance feedback for content-based image retrieval by mining user navigation patterns [J]. IEEE Trans on Knowledge & Data Engineering, 2011, 23 (3): 360-372.

[2] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos [C]// Proc of IEEE International Conference on Computer Vision. [S. l.] : IEEE Computer Society, 2003: 1470.

[3] Ciaccia P, Patella M, Zezula P. M-tree: an efficient access method for similarity search in metric spaces [C]// Proc of International Conference on Very Large Data Bases. [S. l.] : Morgan Kaufmann Publishers Inc, 1997.

[4] Wang Jingdong, Shen Hengtao, Song Jingkuan, *et al.* Hashing for similarity search: a survey [EB//OL]. (2014) . arXiv preprint arXiv: 1408. 2927.

[5] Weiss Y, Torralba A, Fergus R. Spectral hashing [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.] : Curran Associates Inc, 2008: 1753-1760.

[6] Liu Wei, Wang Jun, Ji Rongrong, *et al.* Supervised hashing with kernels [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.] : IEEE Computer Society, 2012: 2074-2081.

[7] Gao Lianli, Song Jingkuan, Zou Fuhao, *et al.* Scalable multimedia retrieval by deep learning hashing with relative similarity learning [J]. Environmental Modelling & Software, 2015, 39 (39): 903-906.

[8] Raginsky M, Lazebnik S. Locality-sensitive binary codes from shift-invariant kernels [C]// Advances in Neural Information Processing Systems. 2009: 1509-1517.

[9] Zhu Xiaofeng, Zhang Lei, Huang Zi. A sparse embedding and least variance encoding approach to hashing [J]. IEEE Trans on Image Processing A Publication of the IEEE Signal Processing Society, 2014, 23 (9): 3737-50.

[10] Shen Fumin, Shen Chunhua, Shi Qinfeng, *et al.* Hashing on nonlinear manifolds [J]. IEEE Trans on Image Processing A Publication of the IEEE Signal Processing Society, 2015, 24 (6): 1839-51.

[11] Shen Xiaobo, Shen Fumin, Sun Quansen, *et al.* Multi-view latent hashing for efficient multimedia Search [C]// Proc of ACM Multimedia. 2015: 831-834.

[12] Liu Li, Yu Mengyang, Shao Ling. Multiview alignment hashing for efficient image search [J]. IEEE Trans on Image Processing A Publication of the IEEE Signal Processing Society, 2015, 24 (3): 956-966.

[13] Zhou Jie, Ding Guiguang, Guo Yuchen. Latent semantic sparse hashing for cross-modal similarity search [C]// Proc of International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014: 415-424.

[14] Ding Guiguang, Guo Yuchen, Zhou Jile. Collective matrix factorization hashing for multimodal data [C]// Proc of Computer Vision and Pattern Recognition. [S. l.] : IEEE Computer Society, 2014: 2083-2090.

[15] Hartigan J A. A K-means clustering algorithm [J]. Appl Stat, 1979, 28 (1): 100-108.

[16] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.] : Curran Associates Inc, 2012: 1097-1105.

[17] Rasiwasia N, Pereira J C, Coviello E, *et al.* A new approach to cross-modal multimedia retrieval [C]// Proc of International Conference on Multimedia. 2010: 251-260.

[18] Huiskes M J, Lew M S. The MIR flickr retrieval evaluation [C]// Proc of ACM International Conference on Multimedia Information Retrieval. 2008: 39-43.

[19] Guillaumin M, Verbeek J, Schmid C. Multimodal semi-supervised learning for image classification [C]// Proc of IEEE Computer Vision and Pattern

Recognition. 2010: 902-909.

[20] Gong Yunchao, Lazebnik S, Gordo A, *et al.* Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (12): 2916.

[21] Yu Xiang, Zhang Shaoting, Liu Bo, *et al.* Large scale medical image search via unsupervised PCA hashing [C]// Proc of IEEE Computer Vision and Pattern Recognition Workshops. 2013: 393-398.

[22] Jin Zhongming, Li Cheng, Lin Yue, *et al.* Density sensitive hashing [J]. IEEE Trans on Cybernetics, 2014, 44 (8): 1362-1371.